Machine Learning Lab Report
Assignment 7

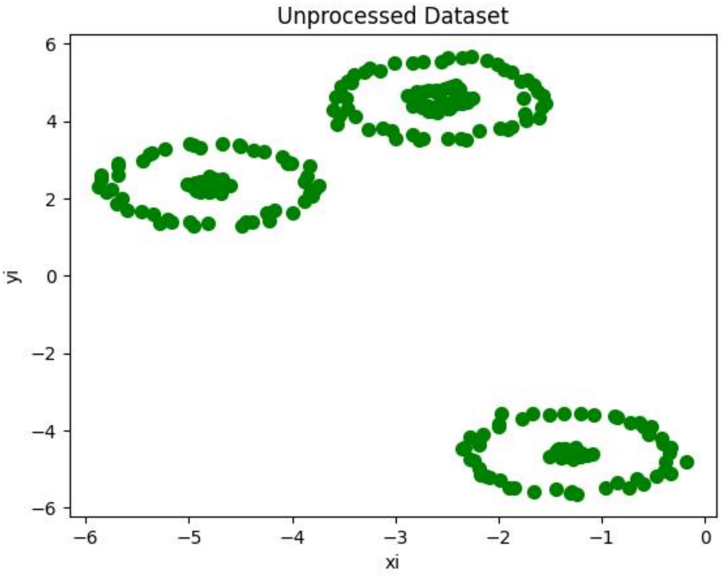
SOUMYA MUKHERJEE
CH24M571
M. Tech in Industrial AI
Trimester -2

Introduction :

This report explores clustering strategies to identify and separate the inner and outer parts of each circle based on the data provided .

Data Exploration :

We observe the data is of size 300 * 2 with 2D data points . No missing or null values were observed . We did an initial scatter plot to visualize the data before clustering .



Clustering :

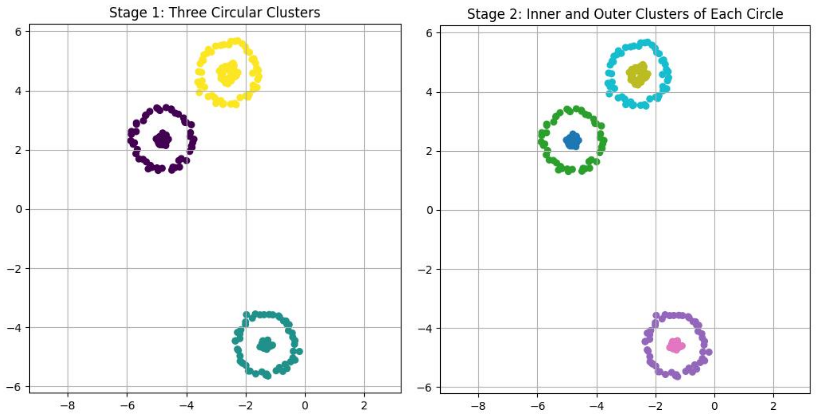
Stage - 1 : K Means is applied with 3 clusters to identify the main circular groups in the data. The resulting clusters are visualized using different colors. A two-stage clustering is performed also using Agglomerative Clustering with 'ward' linkage.

Reasoning for using KMeans Clustering

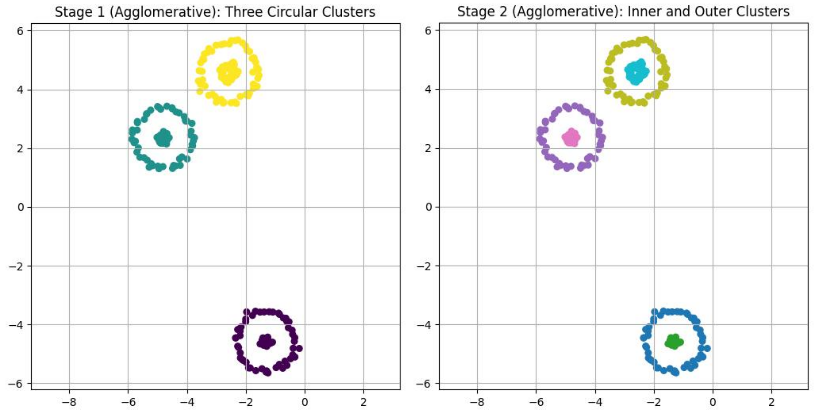
Stage - 2 :

Within each of the 3 clusters from Stage 1, a second KMeans clustering is performed to separate inner and outer points based on their distance to the cluster centroid. This results in a total of 6 clusters. The final clusters are visualized using different colors.

The results are visualized in separate plots for Stage 1 (3 clusters) and Stage 2 (6 clusters).



KMeans Clustering



Agglomerative Clustering

KMeans appears to produce more distinct and well-separated inner/outer clusters compared to Agglomerative Clustering, especially for the smaller circles . The difference is not that huge .

KMeans was chosen for its simplicity and efficiency in this scenario where we know the number of clusters is 3. It's relatively fast and generally provides good results for well-separated, spherical clusters like the circular formations in our dataset (which we can see from the initial plot) . In stage 2 , we're essentially clustering the distances of points from the cluster center. KMeans with 2 clusters effectively separates the points into two groups based on their distance from the centroid (inner and outer).

We implemented Agglomerative Clustering too , but chose to go with K Means since:

> KMeans focuses on minimizing the distance between points and their assigned cluster centers. This tends to produce more spherical and compact clusters, which aligns well with the circular formations in the data.
> Agglomerative Clustering can sometimes be sensitive to the shape and density of clusters, potentially leading to slightly less distinct inner/outer separations in this

case, particularly for the smaller circles that might have less clear density differences between inner and outer parts.

Classification

After obtaining cluster labels from the two-stage KMeans clustering, the data is used for classification.

Data Splitting: The data is split into training and testing sets using a 70-30 ratio with stratified sampling.

Classifiers:

Random Forest: A Random Forest classifier is trained on the training data and evaluated on the test data. The classification report and confusion matrix are displayed.
ExtraTrees: Similarly, an ExtraTrees classifier is trained and evaluated, and its performance metrics are displayed.. ExtraTrees was used since it randomises the tree selection , thereby potentially reducing the variance .
Naive Bayes: A Gaussian Naive Bayes classifier is used, and its performance is evaluated using a classification report and confusion matrix. The confusion matrix is visualized using a heatmap.
Support Vector Machine (SVM): An SVM classifier with an RBF kernel is trained and evaluated, and its performance metrics are displayed.
K-Nearest Neighbors (KNN): A KNN classifier is trained and evaluated. Its decision boundaries are visualized along with the training and testing points.

Classification Reports :

Classifier	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Random Forest	**0.99**	**0.99**	**0.99**	**0.99**
ExtraTrees	**1.00**	**1.00**	**1.00**	**1.00**
Naive Bayes	**0.98**	**0.98**	**0.98**	**0.98**
SVM	**0.79**	**0.64**	**0.59**	**0.64**
KNN	**0.99**	**0.99**	**0.99**	**0.99**

Observations:

ExtraTrees achieved perfect classification with 100% accuracy and perfect precision, recall, and F1-score. This indicates an excellent fit to the data and potentially overfitting if evaluated on more comprehensive datasets.
Random Forest, SVM, and KNN all showed very high performance with accuracy scores of 99%.
Naive Bayes had slightly lower performance compared to other methods with an accuracy of 96%.

Visualization:

The predictions of each classifier on the test set are visualized using scatter plots with different colors representing the predicted classes.

One-Shot Clustering

KMeans is applied directly with 6 clusters to perform one-shot clustering, aiming to identify the 6 inner and outer groups in a single step.
The resulting clusters are visualized using different colors.

Conclusion

The notebook demonstrates different clustering and classification techniques applied to a 2D dataset. Two-stage clustering using KMeans or Agglomerative Clustering effectively identifies the underlying structure of the data. Various classifiers show good performance in predicting the cluster labels, with Random Forest and ExtraTrees showing the best accuracies in the report. One-shot clustering with KMeans provides a simpler approach to achieve similar results. The visualizations provide insights into the clustering and classification results, showcasing the effectiveness of the applied methods.

Due to the presence of less data , Extratrees might have overfitted the data . In such a scenario most tree based algos will give great results .